



Digitální knihovny - včera, dnes a zítra

Martin Svoboda

Státní technická knihovna, Praha

M.Svoboda@stk.cz

Digitálne knižnice, 2. – 3. 10. 2007, Jasná pod Chopkom



Přehled

- definice
- trocha historie
- možné klasifikace
- okruhy otázek dne
- budoucnost?



Definice

- **knihovna** – sbírka užitečného materiálu pro obecné použití - Wikipedia
- **digitální knihovna** sestává z **digitálních sbírek, služeb a infrastruktury** k podpoře celoživotního vzdělávání, výzkumu a odborné komunikace a jejich [dlouhodobého uchovávání] Google define:
- aby mohla být považována za digitální knihovny, musí online sbírku informací řídit a mít k ní přístup **komunita uživatelů** – Wikipedia



Historie

- předchůdci – by-product, born-digital
 - noviny např. The Globe and Mail (Toronto) od 1964
 - referátové služby (CAS: od 1966 fotosazba +prohledávatelná mg.páska, 1980 online via Dialog, 1997 na webu via STN)
- nejstarší - dříve než web – ruční digitalizace
 - Gutenberg (*1971 - Michael Hart, na webu od 1991, texty)
 - Perseus (*1985, na webu od 1995, texty, mapy, obrázky)



Možné klasifikace

- materiál
 - text × audio × multi-media
- vznik
 - born-digital × digitized
- organizace
 - institutional × collaborative
- vyhledávání
 - lokální × federované - broadcast search Z39.50, SRU
- harvested OAI-PMH



Okruhy otázek

- formáty dat
- metadata
- vyhledávání
- dlouhodobé uchování
- ekonomika
- práva duševního vlastnictví
- řízení přístupu a bezpečnost



Formáty dat

- Gutenberg Plain Vanilla ASCII
- texty pdf, TeX,... obrázky jpg, png, djv,...
zvuk mp3,... video avi, ...
- jinými slovy
 - jednoduchost, nezávislost => portabilita
 - ×
 - multimedialita, komfort => závislost



Metadata

- identifikační, popisná, technická, správní
- model OAI
- souvislost se vznikem dat
 - born-digital × digitalizovaná (např. EXIF u fotografií)
- popisná metadata - klasifikace (konsistence mezi indexátory Medline < 49% - Humphrey, 1992)




Vyhledávání⁽¹⁾

- hledání v bohatství velkých knihoven je „jako pěstní klín v ruce uměleckého truhláře“ Vanevar Bush 1945
- v digitální knihovně lze teoreticky hledat daleko přesněji, úplněji
- 1/4 středoškoláků má potíže zvládnout i jediný vyhledávací systém – Borgman, 2001
- Google možná neukáže všechno × OPAC je nepohodlný



Vyhledávání⁽²⁾

- další atributy: podobnost, obálka, žánr, ...
- strategie postupného zpřesňování
- podpora hledajícího
- **klíčový problém**; bez radikálního zlepšení nebude knihovny nikdo potřebovat na nic než na opatřování dat, k nimž se dostane přes 



Dlouhodobé uchování

- e-dokumenty: unikátní, měnící se, mizící,...
 - unikátní identifikátory
- fyzické uchování
 - migrace – emulace – technické muzeum
- uchování odkazů-propojení
 - OAI – ORE ... viz dále
- náklady !!!



Ekonomika

- DL mohou řešit problém distribuce malých nákladů (vědecká publikace často <200 výtisků **v US!!!**, v ČR např. učební texty,...)
- ekonomické (obchodní) modely
 - Open access platí se za publikaci
 - Licensed access platí se za užití
 - vazba na hodnocení výzkumu
- × vyhledávací stroje bezplatně (kde je cena skrytá?)



Práva duševního vlastnictví

- Obhájci tvrdí, že © podporuje vývoj. Skutečně?
 - fonty jsou v Evropě patentovatelné, v US ne, přesto vývoj je především v US
 - Sony vs. Universal: užití videorekordérů **zvýšilo** tržby video průmyslu
 - Umístění dokumentů ILO na web zvedlo tržby z prodeje o 20%
- Alternativní modely:
 - Creative Commons – jasně formulovaná licence pro laika+text pro právníky+text pro počítače
 - šest typů, CC podporují i GNU GPL a GNU LGPL, uznaná v 38 zemích

Řízení přístupu a bezpečnost

- uživatel:
 - autentikace, autorizace
- objekt
 - identita objektu, permanence,...
 - unikátní identifikátor
 - iniciativa OAI: Object Re-Use & Exchange – zavádí Compound Information Objects, které musí „držet pohromadě“





Budoucnost - technicky

- To predict is difficult; especially about the future
(Chinese proverb, A. Einstein, N. Bohr, L.B. Mayer)
- Ale přece jenom pár faktů
 - výkon procesoru $\times 2$ za 18 měsíců
 - kapacita paměti $\times 2$ za 12 měsíců
 - šířka pásma $\times 2$ za 9 měsíců
- 1PetaByte =
 - 100.000 filmů @10GB – tolik ještě není natočeno
 - 100.000.000 obrázků @10MB – stačí na celý život
 - 3000 roků hudby – víc než kolik bylo dosud napsáno či přehráno
- to vše bleskově dostupné

ano, ale jak v tom něco najít?





Co se od nás asi čeká

- z jediného místa (ne 24 portálů), všechny druhy dokumentů k dané otázce, uspořádání podle relevance, možnost určení
- intuitivní ovládání, ne školení
- pro-aktivní, nabízející přístup
- nadstavby: My..., VLE



Techniky

- ontologie, našeptávače,
- analýzy protokolů vyhledávání: „ti, kdo se dívali na tohle, také prohlíželi ...“
- individualizace, „znalost zákazníka“
- participace - kolaborace - social computing, tagging
- staronové metody
 - facetové hledání (e.g. flamenco)



Strategie

- přístup
 - centralizace – jediný stroj (srv Bielefeld BASE)
- zdroje:
 - český – česko-slovenský JSTOR?
 - navnadit velké vydavatele časopisů k dostupnosti online?
- netříštit síly: na společných projektech se dá získat dost slávy pro všechny – je nás jen 15 milionů



Děkuji za pozornost

M. Svoboda@stk.cz



Nobelovu cenu

pro otce Webu

Sira Tima

Berners-Lee!